International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-4, Issue-1, January 2016

A Survey on Feature Selection in Data Mining

Dr.M.Chidambaram, R.Umasundari

Abstract— **Feature Selection is a fundamental problem in machine learning and data mining . Feature Selection is an effective way for reducing dimensionality, removing irrelevant data increasing learning accuracy. Feature Selection is the process of identifying a subset of the most useful features that produce compatible results as the original entire set of features .A Feature Selection techniques may be evaluated from both efficiency and effectiveness point of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of the subset of features. Feature Selection is different from dimensionality reduction. Both methods search for to reduce the number of attributes in the dataset. But dimensionality reduction method creating new combination of attributes. Feature Selection methods include and exclude attributes present in the data without change. The central assumption when using a Feature Selection technique is that the data contains many redundant or irrelevant features. This paper actually a survey on various technique of feature selection and its advantages disadvantages.**

Index Terms— **Data mining, Feature selection, clustering, Relief ,Minimum spanning tree**

I. INTRODUCTION

Data mining is defined as extracting the information from high dimensional data. In order to effectively extract the information from huge amount of data bases. Data mining algorithm must be efficient and scalable. Many feature selection methods have been proposed and studied for machine learning applications. Cluster analysis serve as a tool to gain insight into the distribution of data to observe characteristics of each cluster. we first partition the set of data into groups based on data similarity and then assign the labels to the groups. Once all these processes are over we would be able to use this information in many application such as fraud detection ,market analysis ,production control science exploration etc.

Feature Selection an important role in the data mining process .This paper made a survey on various existing Feature Selection technique used. Filters use general characteristics of the training data to evaluate attributes and operate independently of any learning algorithm. wrappers evaluate attributes by using accuracy estimates provided by the actual target learning algorithm .The wrapper model is computationally expensive and the filter model is usually a good choice when the number of features becomes very large .

Manuscript received January 09, 2016.

Dr.M.Chidambaram, Assistant professor in Computer Science Department Rajah Serfoji Government college (Autonomous) Thanjavur.

R.Umasundari Research Scholar in Computer Science, Bharathidasan University Constituent college for women, Orathanadu

II. FEATURE SELECTION FOR HIGH DIMENSIONAL GENOMIC MICROARRAY DATA

This paper produce reference [1],[2] a new framework of Feature Selection as a pre-processing step to machine learning, has been efficient in reducing dimensionality, removing irrelevant data ,increasing learning accuracy and improving comprehensibility. The recent increase of dimensionality of data poses a severe challenge to many existing Feature Selection methods with respect to efficiency and effectiveness. In the first step reference[8] ,features are divided into clusters and second step the most representative feature is strongly related to target classes is selected from each cluster to form a subset of features .In this goal paper introduce a novel concept predominant correlation, and propose a energy Efficiency clustering filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. The efficiency and effectiveness of our method is demonstrated through extensive comparisons with other methods using real world data of high dimensionality .Its advantage over the traditional framework of subset evaluate removing irrelevant analysis.

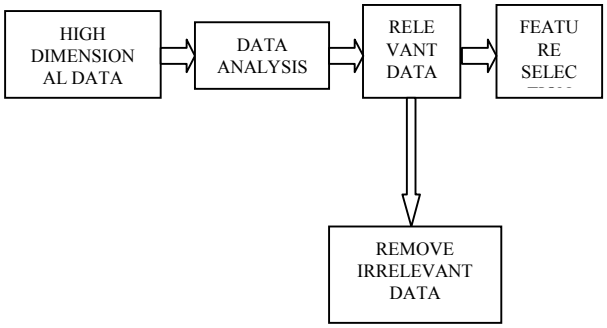

Figure 1: Framework of feature selection

III. EXTENDED RELIEF ALGORITHMS IN INSTANCE BASED FEATURE FILTERING

The paper[5] presents Relief Algorithms and their use in instance based feature filtering for document feature selection .The Relief algorithm are used image data, microarray data ,text data filter the features.

The Relief algorithm are general and successful feature estimators that detect conditional dependencies of features

between instances ,and are applied in the pre processing step for document classification and regression .many kinds of extended Relief algorithm have been suggested as solutions to problem of redundancy, irrelevant and noisy feature as well as Relief algorithm limitations in high datasets. These algorithm are used to removing irrelevant feature reduced dataset, and extended reduced searching time and also improve the performance and security. we suggest new extended Relief algorithm to solve those of quality of features from instances and classified datasets.

IV. A ENERGY EFFICIENCY CLUSTERING FOR HIGH DIMENSIONAL DATA

In the high dimensional data set having Feature Selection[14] involves identifying a subset of the most useful feature produce compatible result as the original entire set of features. A feature may be evaluated from the both efficiency concerns the time required to find subset of features .in this goal paper introduce a subset of good features with respect to the target concepts feature subset selection is an effective way for reduce dimensionality removing irrelevant data , increasing learning accuracy and improving result comprehensibility many feature subset selection methods have been proposed and studied machine learning application. The filter methods are independent of learning algorithm good generality..their computational complexity low but the accuracy of the learning algorithm is not guaranteed .the minimum spanning tree based clustering algorithms
,because they do not assume that data points are grouped around canter or separated by a regular geometric curve and have been widely used in practice .based on the minimum spanning tree method we Propose a method energy efficiency clustering based feature selection.

V. FEATURE SELECTION BASED ON A NEW DEPENDENCY MEASURE

Feature selection is a process commonly used in machine learning, where in subsets of the features available from the data are selected for application of a learning algorithm. Features selection is effective in reducing dimensionality, removing irrelevant learning accuracy and efficiency. In this paper[2], We propose a new information distance to measure the relevancy of two features.

VI. FEATURE SELECTION TECHNIQUES IN EDUCATIONAL DATA MINING

Educational data mining is a new research area and the fundamental nature of data mining concepts are used in the educational field for the purpose of extracting useful information on the behaviors of students in the learning process. As the feature selection the analytical accuracy of any performance model, it is essential to study elaborately the effectiveness of student performance model in connection with feature selection techniques. The main objective of feature selection is to choose a subset of input variables by reducing features, which are irrelevant or of no analytical information. Feature selection in supervised learning has a main goal of finding a feature subset that produces higher classification accuracy

VII. CONCLUSION

Feature Selection is a term commonly used in data mining to describe the tools and techniques available for reducing data to a manageable size for processing and Data analyze. This paper explains about the data mining functionalities and also about the Feature subset Selection. Feature selection technique has wide variety of application in data mining ,text mining , digital image processing etc. Feature Selection Technique and its advantages as well as disadvantages are depicted in this paper. Feature selection via explaining training samples using concepts generalized from existing features and knowledge. In many real-world applications, the same trend might be caused by distinct reasons. Feature Selection Technique and its advantages as well as disadvantages are depicted in this paper.

REFERENCES

[1] Feature selection for high-dimensional genomic microarray data Feature selection for high-dimensional genomic microarray data, Eric P.Xing, Michael I.Jordan, Proceeding of the Eighteenth International Conference on Machine Learning Pages 601-608 .

[2]Eric P.Xing Feature selection based on a New Dependency measure,2008 Fifth international conference on FUZZY Systems and knowledge discovery,.

[3] Zhao Z. and Liu H. (2009), 'Searching for Interacting Features in Subset Selection', Journal of Intelligent Data Analysis, 13(2), pp 207-228, 2009.

[4] Xing E.,Jordan M. And Karp R., Features selection for high –dimensional genomic microarray data,2001.

[5] Park H. and Kwon H., Extended Relief Algorithms in Instance-Based Feature Filtering, In Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT2007), pp 123-128, 2007.

[6] Sha C., Qui x and Zhou A., Feature selection based on a New Dependency measure,2008 Fifth international conference on FUZZY Systems and knowledge discovery,.

[7] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279-305, 1994

[8] Qinbao Song, Jingjie Ni and Guangtao Wang, A Fast Clustering Based Feature Subset Selection Algorithm for
High Dimensional Data, IEEE Transactions on Knowledge and Data Engineering vol:25 no:1 year 2013.

[9] Heum Park,Hyuk-Chul kwon Extended Relief Algorithms in Instance-Based Feature Filtering.22-24 Aug.2007.

[10] Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure based on relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109,2004.

[11]E. Baker, International Encyclopedia of Education (3rd edition), Oxford, UK: Elsevier

[12] P. Mitra, C. A. Murthy and S. K. Pal. "Unsupervised feature selection using feature similarity," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301–312,2002.

[13] Miller, "Subset Selection in Regression," Chapman & Hall/CRC (2nd Ed.), 2002.13.H. Almuallim and T. G. Dietterich. "Learning boolean concepts in the presence of many irrelevant features," Artificial Intelligence,vol. 69, no. 1-2, pp. 279–305, 1994.

[14] A Energy Efficiency Clustering for High Dimensional Data M.Ramzanfathimakani [1], R. Bharathkumar [2] Pristuniversity,kumbokanam.